Stephanie Mecham

Section 6

BIOSTAT 522 HW #3 (Due **Thurs, Feb 1**, 2018, in class)

We will continue to use ***surgery*** dataset for this homework. As before, whenever you are asked to test for a hypothesis, write the null and alternative hypotheses, give the formula for the test statistic (in symbols), the numerical value of the test statistic, and the p value and draw a conclusion.

1. The variables **depression** and **anxiety** are the self-reported depression and anxiety, respectively, measured on the same day using a scale that can give values ranging between 0 (corresponds to min possible level) to 10 (max).

1-A) Fit a regression model with **depression** predicting **anxiety** and conduct an **F-test** of H0: β1 = 0 vs. HA: β1 ≠ 0. For the numerical value of the test statistic, please make sure to show how the F statistic was calculated based on values given in the ANOVA table. Also show how the **F test statistic** and the **R2**are related using the numerical values you obtained.

F-value for this regression model is 316.21. This was calculated from the ANOVA table as the (Mean Square model)/ (Sum of Squareserror / DFerror) = 828.33910 / (518.685/198) = 316.21. The F-statistic and R2  are related by the formula F=R2 /(1-R2 ) x (dfE /dfM) which in our model is (0.6149/(1-0.6149)) x (198/1), which evaluates to 316.15, which is our F-statistic.

1-B) For any subject *i*, the difference between the fitted value of Y and the overall mean of Y, [yihat - ybar] is a multiple of the difference between that subject’s predictor and predictor mean, [xi - xbar]. Or, for individuals with a covariate value x0, [ mu (x0) - ybar ], i.e., the difference between their predicted mean of Y and the mean of Y, is a multiple of [x0 - xbar], i.e., the difference between those individuals’ predictor and predictor mean. **What is the multiplier in this problem?**

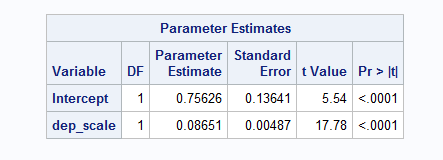
(xi - xbar)\*M = (yhat-ybar) → where M stands for multiplier.

We know that yhat = βhat0 +( β1\*xi) and βhat0 = ybar - (βhat1 \* xbar).

Substituting the βhat0 equation into the yhat equation, we get yhat= ybar - (βhat1 \* xbar) + (βhat1 \* xi). We can rearrange this equation to get yhat - ybar = - (βhat1 \* xbar) + (βhat1 \* xi), which is equivalent to yhat - ybar = βhat1 (xi - xbar). This is our original equation, which shows that βhat1 is our multiplier (0.8651 in our case).

1-C) **Answer this problem using the output from #1-A.** Suppose other researchers often use depression measured in a 0–100 scale (rather than 0 to 10). For this reason, you’d like your slope estimate to be interpreted comparably to other studies. For the model in #1-A, what would be the slope estimate of depression if it were recorded on a 0 to 100 scale?

It would be β1 / 10 = 0.86512 / 10 = 0.0865, because β1 by definition represents the change in the dependent variable for every one unit increase in the independent variable, and we just stretched out our independent variable so that each unit increase from our original scale represents ten unit increases in our new scale. This will cause our β1 to decrease by a factor of 10 because there is less change per unit increase than our original.

1-D) Verify your answer to #1-C by modeling **anxiety** using an appropriately **rescaled depression** so that it corresponds to a 0-100 scale. Submit your SAS output. Write your equation for each model in #1-A and #1-D using the parameter estimates. Interpret intercepts and slopes from each model, and comment briefly about whether the parameter estimates between the two models make sense. 

This output verifies my answer to 1-C.

1-A: y= 0.75626 + 0.86512x, where x is self-reported depression score and y is self-reported anxiety score.

1-D: y= 0.75626 + 0.0865x, where x is self-reported depression score and y is self-reported anxiety score.

The intercepts of 0.75626 represent the estimated self-reported anxiety score for individuals who ranked a depression score of 0. The slope from part A represents that for every unit increase in depression ranking, the anxiety score is expected to increase by 0.86512 units. The slope from part D represents that for every unit increase in depression ranking, the anxiety score is expected to increase by 0.0865 units. The slopes are different because the per-unit change is much higher in the 1-10 scale than in the 1-100 scale.

1-F) If the model fits well, there should be a good correlation between the predicted responses (’s) and observed (yi’s). Obtain corr(, yi) using only the information in the PROC REG output of the model fit for problem #**1-D**.

The R2 value of 0.6149 is a representation of the correlation between predicted and observed values. Because this value is pretty close to 1, we can say that the model fits the data well.

1-G) Suppose for the data used in fitting the model in #**1-A,** *pain* is another variable collected for all study participants. It seems natural that *pain*might be related to anxiety level. We therefore consider the following model: E(Anxiety) = β0 + β1 pain

What is SSY based on this model?

SSY= SSE + SSM = 1028.24599 + 361.98969 = 1390.24

2. Now we will consider variables **depression, anxiety** and **trait**. **Trait** measures the trait anxiety (ranging from 0 to 100). Consider the following two models:

**Model 1**: Relationship of **anxiety** to **depression**

**Model 2**: Relationship of **anxiety** to **depression** and **trait**

If you recall, including multiple models in one PROC REG statement will fit all models using only complete data (non-missing) for all variables included in every model. On the other hand, if you fit each model with separate PROC REG, each will use complete data for the variables included in the model (both Y and X’s). In this problem, you will see that the two approaches can give you different estimates because of the missing data.

2-A) Fill the following table using results from the two models fit with one PROC REG statement (left) and fit with two separate PROC REG statements (right). Rows 4 to 6 will be filled with the parameter estimate and p-value in the parenthesis. Note that some cells should remain blank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **One PROG REG** | |  | **SEPARATE PROG REGs** | |
|  | Model 1 | Model 2 |  | Model 1 | Model 2 |
| N | 186 | 186 |  | 200 | 186 |
| Intercept | 0.75048 (<.0001) | 0.84971 (0.0005) |  | 0.75626 (<.0001) | 0.84971 (0.0005) |
| Depression | 0.87821 (<.0001) | 0.88562  (<.0001) |  | 0.86512 (<.0001) | 0.88562 (<.0001) |
| Trait |  | -0.00373 (0.6092) |  |  | -0.00373 (0.6092) |
| (Root MSE) | 1.58213 | 1.58531 |  | 1.61853 | 1.58531 |
| R2 | 0.6394 | 0.6399 |  | 0.6149 | 0.6399 |
| Ra2 | 0.6375 | 0.6360 |  | 0.6130 | 0.6360 |

2-B) What do you think is the advantage (other than shorter SAS codes) of comparing models 1 and 2 using one PROC REG vs. two separate PROC REG statements?

It is useful when you want to be consistently measuring the same people across variables. Using separate proc regs will result in a different assemblage of people from your sample contributing to the analysis, which could introduce inconsistency between the separate proc reg’s.

2-C) Consider Model 3 where you now add a variable measuring **income** to Model 2. Without fitting the model and assuming no one has missing income, describe in a couple of sentences what you expect for each of SSM, SSE and SSY from Model 3 compared with the SSM, SSE and SSY from Model 2.

The addition of another variable to the model would likely decrease the SSE (taking into account an additional variable will account for some of the unexpected spread of the means and errors, whereas in one-variable models you assume everything that deviates from the model prediction is purely random scatter). Expecting income to have an association with anxiety, I would expect an increase in the SSM. Overall I would expect an approximately equal SSY value.

3. For this problem, you have to fit 4 regression models for **anxiety**. But first fit a model with **depression**, **trait**, and **etoh** as predictors**.** **Etoh** assesses alcohol dependence: 0=not dependent, 1=mildly dependent, 2=dependent, and 3=highly dependent.

3-A) Write the equation for the model fit in three ways:

1. with **anxietyi** on the left side of the equal sign, and using symbols (i.e., betas),
2. with the mean, **E(anxietyi | depressioni, etohi, traiti)**, on the left side of the equal sign, and
3. with the parameter estimates.
4. Anxietyi = β0 + β1(depression)+ β2(etoh)+ β3(trait)+ εi
5. E(anxietyi | depressioni, etohi, traiti) = β0 + β1(depressioni)+ β2(etohi)+ β3(traiti)
6. Ŷi = 0.88450 + 0.87148(depressioni) - 0.07881(etohi) - 0.00236 (traiti)

3-B) Interpret the regression parameter estimates for **etoh** and intercept.

The parameter estimate for etoh is -0.07881, which represents that for each unit increase in alcohol dependence level, the predicted anxiety score will decrease by 0.07881. This only results in a p-value of 0.4929 though, so the association is not strong enough to reject the null. The parameter estimate for the intercept is 0.88450, which indicates that for individuals with a score of 0 for depression, a score of 0 for anxiety trait, and a score of 0 on the alcohol dependency scale, we can estimate their self-reported anxiety score to be 0.88450.

3-C) Interpret the R2, and verify the R2 and the adjusted R2 given for the model using the values given in the ANOVA table.

The R2 is 0.6259, which shows that the model is a good fit of the data because it is relatively close to 1. We can verify this by the equation R2 = 1-(SSE/SST), which in our model evaluates to 1-(452.79770)/(452.79770+757.46469) = 0.62587, which is almost exactly what we got from our output. To verify the adjusted R2, we can use the formula R2adj = 1- [(1-R2)(n-1)/(n-k-1)] where n is number of observations used and k is number of independent regressors. In our model, this evaluates to [(1-0.62587)(184-1)/(184-3-1)] = 0.6196, which is exactly what we got in our ANOVA table.

3-D) Based on the model fit, what is the **predicted anxiety** for a person with mild alcohol dependence, trait of 45 and depression of 2?

Using our model, anxiety = 0.88450 + 0.87148(depressioni) - 0.07881(etohi) - 0.00236 (traiti), we can fill in our equation to get 0.88450 + 0.87148(2) - 0.07881(1) - 0.00236 (45). This evaluates to a predicted anxiety score of 2.44.

3-E) Based on the model, find the **difference in anxiety score** corresponding to an increase in depression from 3 to 4 for a person with no drinking dependence and trait of 50.

Model: Anxiety = 0.88450 + 0.87148(depressioni) - 0.07881(etohi) - 0.00236 (traiti)

Model @ depression value of 3:

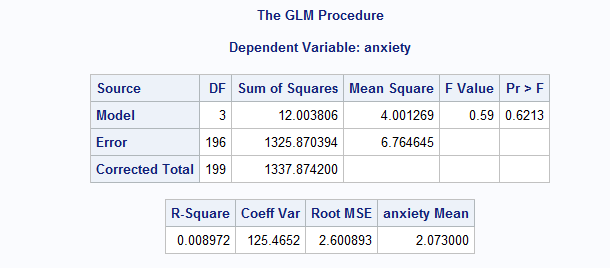
Anxiety = 0.88450 + 0.87148(3) - 0.07881(0) - 0.00236 (50) = 3.380

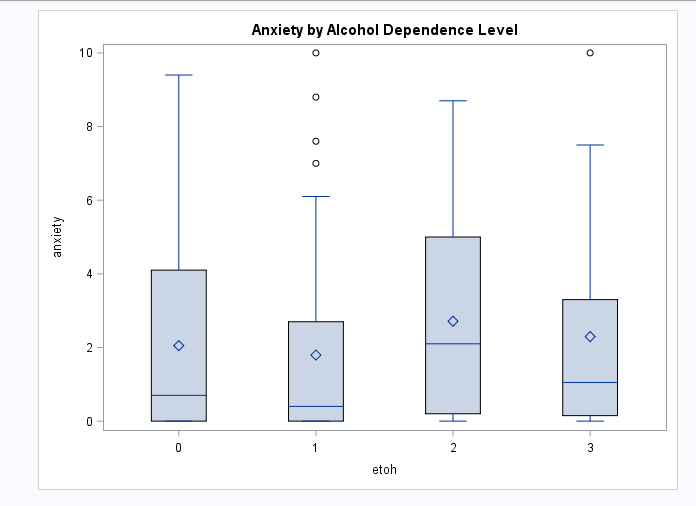
Model @ depression value of 4:

Anxiety = 0.88450 + 0.87148(4) - 0.07881(0) - 0.00236 (50) = 4.252

Difference in anxiety score from depression score of 3 to depression score of 4: 4.252-3.380 = 0.872 (which makes sense, because that is the value of our slope for the depression variable).

3-F) Using SAS, assess the crude relationships between **anxiety** and **etoh** (4 level variable) and obtain appropriate summary statistics (e.g., mean **anxiety** at each **etoh** level) and provide one appropriate graph showing the relationship.





Mean anxiety at etoh level 0: 2.050

Mean anxiety at etoh level 1: 1.794

Mean anxiety at etoh level 2: 2.712

Mean anxiety at etoh level 3: 2.296

3-G) Using a **simple linear regression**, assess the unadjusted relationships between **anxiety** and **etoh** (4-level variable). Find the answer to **#3-D** based on this new model (**predicted anxiety** for a person with mild alcohol dependence, trait of 45 and depression of 2):

Model: Anxiety = 0.88450 + 0.87148(depressioni) - 0.2561(etohi) - 0.00236 (traiti)

= 2.05 – 0.2561

= 2.3061

3-H) Now fit a model with same predictors (**depression**, **trait**, and **etoh)**, but instead of **etoh,** includea variable indicating for **any level of alcohol dependence**, i.e., etoh=1, 2, or 3 vs. 0**.** Youneed to generate such a variable from **etoh**. Write the model with the parameter estimates, and find the answer to **#3-D** based on this new model. In one sentence, contrast your answers to #3-D.

Model: Anxiety = 0.8994 + 0.8710(depressioni) - 0.1694(dependency) - 0.0023 (traiti)

Predicted anxietyfor a person with mild alcohol dependence, trait of 45 and depression of 2:

Anxiety = 0.8994 + 0.8710(2) - 0.1694(1) - 0.0023 (45) = 2.3685. This value is a little bit lower than our value we got in 3-D (2.44 - 2.3685 = 0.0715)

.

3-I) Now create dummy variables based on **etoh** for each level. Fit a model evaluating the relationship between **anxiety** and **depression**, controlling for **trait** and **the different alcohol dependence levels.** Use never drinker (**etoh** = 0) as the referent category. Write your equation for the above model using the parameter estimates you found. Interpret each of the alcohol dependence coefficients (parameter estimates), and contrast them to the interpretation of etoh of #3-B.

Model:

Anxiety = 0.91800 + 0.87086(depression) - 0.00299(trait) - 0.18407(mild dependence) + 0.07406(dependence) - 0.29588 (high dependence)

The parameter estimate for mild alcohol dependence is slightly negative (-0.184), for moderate dependence is slightly positive but very close to zero (0.074), and for high dependence is negative, even more so than the estimate for mild dependence (-0.297). We can interpret this as there being a slightly positive association between moderate alcohol dependence and anxiety score, and a negative association between mild or high alcohol dependencies and anxiety scores. These are much more specific than the catch-all value from 3B, which showed a slightly negative association between alcohol dependency and anxiety score (parameter estimate of -0.07881).

3-J) Interpret o (intercept) under the three models described:

1. model fit for #3-A, with **etoh,**
2. model fit for #3-H, with an indicator for any level of alcohol dependence,
3. model fit for #3-I, with indicators for the different alcohol dependence levels.
4. The intercept of the model from 3-A is 0.88450. This represents the predicted anxiety score of an individual who had a 0 depression score, no alcohol dependence, and no anxiety trait.
5. The intercept of the model from 3-H was 0.8994. This represents the predicted anxiety score of an individual who had a 0 depression score, no form of alcohol dependence, and no anxiety trait. It is different than the intercept from part 1 because in part 1, having etoh was multi-leveled and in part 2 it is binary.
6. The intercept of the model from 3-I was 0.91800. This represents the predicted anxiety score of an individual who had a 0 depression score, no alcohol dependence of any kind, and no anxiety trait. This is slightly different from the intercepts from the other parts because in this case, each dependency type has been separated out and the effects have been analyzed individually, rather than being lumped together in a dichotomous or ordinal variable like in parts 1 and 2.

3-K) Estimate the mean difference in anxiety associated with mild alcohol dependence vs. no dependence, for 30 years old with one comorbidity, depression score of 3 and trait score of 30 based on the three different models described in #3-J.

Model from 3-A:

Mild dependence:

Anxiety = 0.88450 + 0.87148(depressioni) - 0.07881(etohi) - 0.00236 (traiti)

Anxiety = 0.88450 + 0.87148(3) - 0.07881(1) - 0.00236 (30)

Anxiety = 3.349

No dependence:

Anxiety = 0.88450 + 0.87148(depressioni) - 0.07881(etohi) - 0.00236 (traiti)

Anxiety = 0.88450 + 0.87148(3) - 0.07881(0) - 0.00236 (30)

Anxiety =3.428

Difference: -.07914

Model from 3-H:

Mild dependence:

Anxiety = 0.8994 + 0.8710(depressioni) - 0.1694(dependency) - 0.0023 (traiti)

Anxiety = 0.8994 + 0.8710(3) - 0.1694(1) - 0.0023 (30)

Anxiety = 3.274

No dependence:

Anxiety = 0.8994 + 0.8710(depressioni) - 0.1694(dependency) - 0.0023 (traiti)

Anxiety = 0.8994 + 0.8710(3) - 0.1694(0) - 0.0023 (30)

Anxiety = 3.4434

Difference: -0.1694

Model from 3-I:

Mild dependence:

Anxiety = 0.91800 + 0.87086(depression) - 0.00299(trait) - 0.18407(mild dependence) + 0.07406(dependence) - 0.29588 (high dependence)

Anxiety = 0.91800 + 0.87086(3) - 0.00299(30) - 0.18407(1) + 0.07406(0) - 0.29588 (0)

Anxiety = 3.257

No dependence:

Anxiety = 0.91800 + 0.87086(depression) - 0.00299(trait) - 0.18407(mild dependence) + 0.07406(dependence) - 0.29588 (high dependence)

Anxiety = 0.91800 + 0.87086(3) - 0.00299(30) - 0.18407(0) + 0.07406(0) - 0.29588 (0)

Anxiety = 3.441

Difference: -0.184

-( .07914 + 0.1694 + 0.184) / 3 =-0.144 is the mean difference in anxiety.

3-L) Which of the **five models considered so far** (three models described under #3-J and Models 1 and 2 of problem #2) appears to be the best model based on the appropriate criterion you learned for selecting a model?

I would choose Model 2 as the best based on its high adjusted R2 relative to the other models.

Adjusted R2:

Model 1 → 0.6130

**Model 2 → 0.6360**

Model 3-A → 0.6196

Model 3-H → 0.6197

Model 3-I → 0.6166

**Code:**

**\*** BIOSTATS 522 HW 3 | STEPHANIE MECHAM | SECTION 6 |;

libname lab "C:\Users\smecham\Desktop\lab";

\*Question 1-A: Regression Model of Depression Predicting Anxiety;

proc reg data=lab.surgery;

model anxiety = depression;

run;

\*Question 1-D: Re-scaling Depression Variable;

data lab.surgery\_rescale;

set lab.surgery;

dep\_scale = depression\*10;

run;

\*Regression Model of Re-scaled Depression Predicting Anxiety;

proc reg data=lab.surgery\_rescale;

model anxiety = dep\_scale;

run;

\*Question 1-G: Regression Model of Pain Predicting Anxiety

proc reg data=lab.surgery;

model anxiety = pain;

run;

\*Question 2: Comparing Combined vs Isolated Regression Models;

\*Combined Proc Regs;

proc reg data=lab.surgery;

model anxiety = depression;

model anxiety = depression trait;

run;

\*Isolated Proc Regs;

\*Model One;

proc reg data=lab.surgery;

model anxiety=depression;

run;

\*Model Two;

proc reg data=lab.surgery;

model anxiety=depression trait;

run;

\*Question 3: Regression Model with Depression, Trait, and EtOH as Predictors;

proc reg data=lab.surgery;

model anxiety = depression trait etoh;

run;

\*Question 3 Part F;

proc glm data=lab.surgery plots=all;

class etoh (ref="0");

model anxiety=etoh;

run;

proc sgplot data=lab.surgery;

vbox anxiety / category=etoh;

title "Anxiety by Alcohol Dependence Level";

run;

proc means data=lab.surgery;

class etoh;

run;

\*Question 3 Part G;

proc reg data=lab.surgery;

model anxiety = etoh;

run;

proc genmod data=lab.surgery plots=all;

class etoh (ref="0");

model anxiety = etoh;

run;

quit;

\*Question 3 Part H: Creating a Binary Alcohol Dependence Variable;

data lab.surgery;

set lab.surgery;

if etoh = . then dep=.;

else if etoh > 0 then dep=1;

else dep=0;

run;

\*Model;

proc reg data=lab.surgery;

model anxiety = depression trait dep;

run;

\*Question 3 Part I: Creating Dummy Variables for Alcohol Dependence;

data lab.surgery;

set lab.surgery;

if etoh=. then no\_dep=.;

else if etoh=0 then no\_dep=1;

else no\_dep=0;

if etoh=. then mild\_dep=.;

else if etoh=1 then mild\_dep=1;

else mild\_dep=0;

if etoh=. then dep=.;

else if etoh=2 then dep=1;

else dep=0;

if etoh=. then high\_dep=.;

else if etoh=3 then high\_dep=1;

else high\_dep=0;

run;

\*Model;

proc reg data= lab.surgery;

model anxiety = depression trait mild\_dep dep high\_dep;

run;

\*End of code;